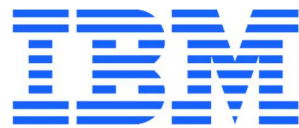


IBM solution in AI Sweden lab



Content:

IBM AI lab solution capabilities..... 3

 IBM Fusion HCI 3

 IBM Fusion in AI Sweden lab – the specifications 6

 Watsonx data & AI platform..... 6

 AI building blocks to the future 9

 AI capabilities are growing rapidly..... 9

 IBM Model strategy..... 10

 AI Governance 11

 IBM Cloud 13

 AI Infrastructure 15

 IBM Security Compliance Center and Workload Protection..... 16

 IBM Client Engineering..... 18

DDO use case mappings 18

DDO – Best practices for operationalization of AI in Sweden 19

IBM watsonx presentation and demo 19

IBM Contacts & Contribution..... 20



IBM AI lab solution capabilities

In this document we have described the existing solution capabilities installed at IBM-supplied Fusion machine the AI Sweden lab in Gothenburg. The locally installed AI services in combination with public and private cloud services are mentioned on a high level, with more detailed information in supported appendices.

In a separate document (DDO project IBM contribution) we have also mapped the candidate AI lab solution capabilities/services from IBM to each of the DDO use cases. The purpose is to inform how the AI lab solution capabilities can fit and meet the requirements of each individual DDO use case.

The following IBM solutions are installed at AI Sweden lab in Gothenburg:

- **IBM Fusion HCI**, hardware platform designed for Red Hat OpenShift workloads, **watsonx** data & AI platform, and OpenShift virtualization
- **IBM watsonx**, the software platform for end-to-end data & AI capabilities, a selected set of components from the broad range of services **watsonx** provides are installed
- **IBM Cloud** platform account with services & AI infrastructure capabilities

IBM Fusion HCI

Fusion HCI (Hyper-Converged Infrastructure) offers three powerful capabilities that address critical enterprise needs. First, it functions as an “Data & AI Platform in a box”, enabling teams to quickly provision the infrastructure needed for generative AI fine-tuning and inferencing—significantly shortening deployment cycles. Second, it serves as a VM Migration Platform, helping organizations transition legacy virtual machines to a more modern, cost-efficient environment.

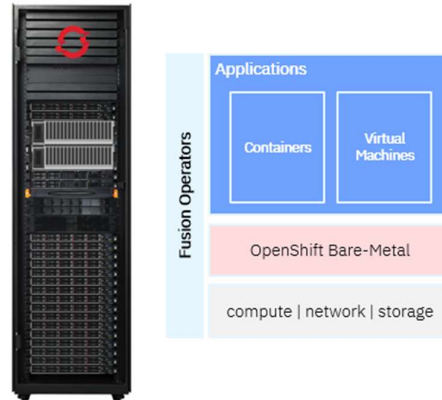


Figure 1 - IBM Fusion HCI

Finally, Fusion HCI acts as a Cluster Vending Machine, giving developers on-demand access to Kubernetes clusters for rapid experimentation and innovation.

These capabilities directly translate into measurable business outcomes. The data & AI Platform feature empowers data science teams to bring AI insights to market faster, improving competitive advantage and enabling data-driven decision making.

By streamlining the VM migration process, Fusion HCI helps organizations reduce cost, free up IT budgets, and focus resources on strategic initiatives rather than maintenance.

And through its Cluster Vending Machine functionality, Fusion HCI supercharges developer productivity, enabling faster application cycles and continuous delivery across development, test, and production environments.

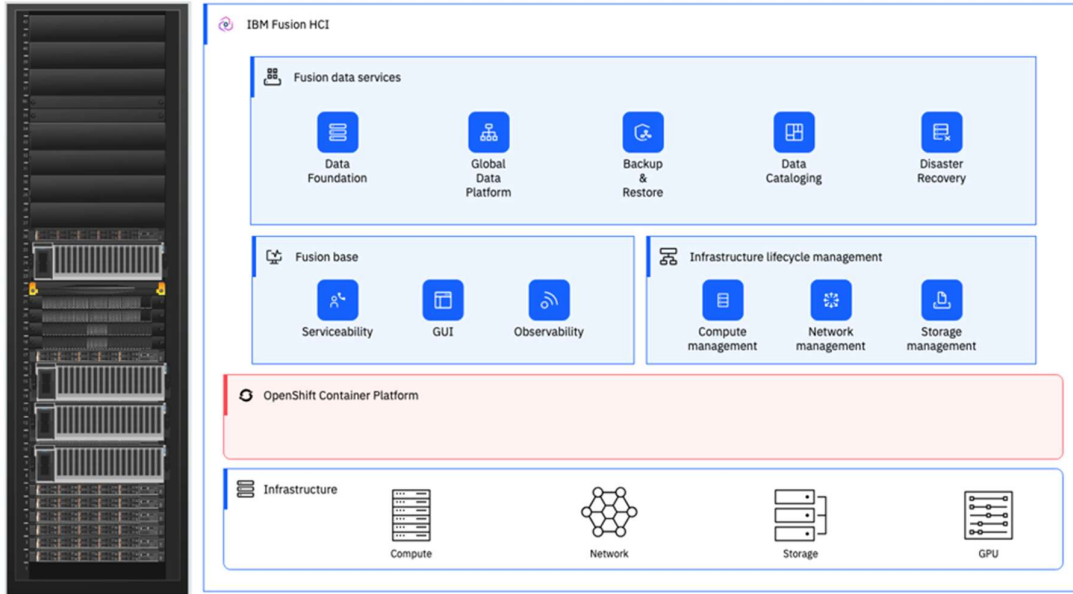


Figure 2 - Fusion HCI with Red Hat OpenShift and IBM watsonx

Can Fusion HCI support production grade workloads? Absolutely yes. Fusion is designed from the ground up to support enterprise-grade, mission-critical workloads, including those in production environments. By uniting bare-metal OpenShift, robust storage options, HA/DR services, and full-stack operator-driven automation, Fusion provides the reliability, security, and performance your business needs to run production clusters with confidence. From automated updates and lifecycle management to backup and disaster recovery capabilities, each component has been engineered to meet the demands of modern, high-volume, and high-availability applications.

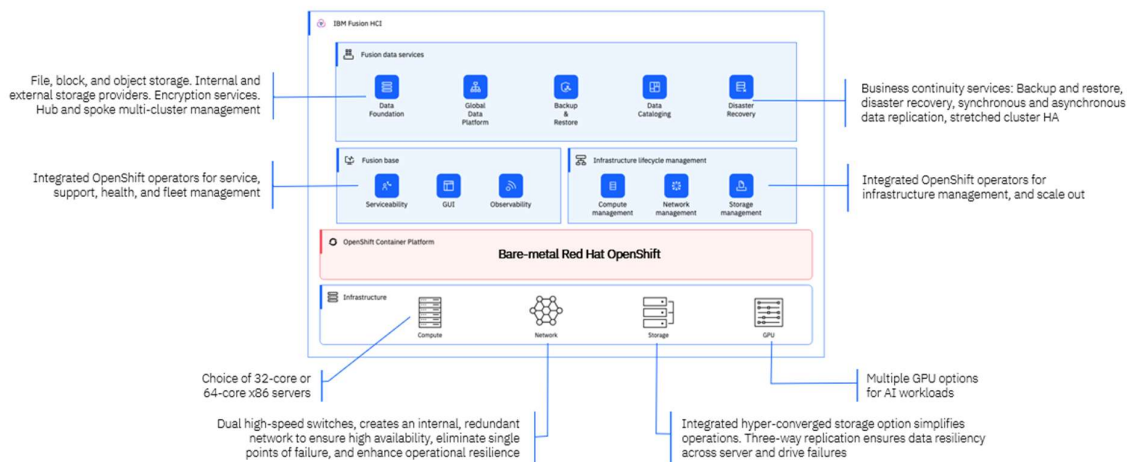


Figure 3 - Fusion HCI Architecture Software Stack

Fusion HCI is already in production with IBM's watsonx data & AI platform at clients' datacenters in Sweden, delivering business critical services today.

See further, IBM Storage Fusion Product Guide: chapter xx

See further, Accelerating IBM watsonx.data with IBM Fusion HCI:

<https://www.redbooks.ibm.com/redpapers/pdfs/redp5720.pdf>

IBM Fusion in AI Sweden lab – the specifications

The following resources and components are installed in the AI Sweden lab:

- 3 Master nodes
- 3 Worker nodes á 32 cores
- 1 GPU worker node á 96 cores
- 4 Nvidia L40S GPU
- IBM Fusion SW (including Fusion Data Foundation and Fusion Backup services)
- RedHat OpenShift Container Platform
- IBM watsonx.ai, (GenAI & machine learning)
- IBM watsonx.data, (Lakehouse & Milvus vector database)
- IBM watsonx.governance, (AI governance)
- IBM Watson Knowledge Catalog, (Data governance, data catalog, data quality)
- IBM Db2, (database)
- IBM Granite 3-2-8B-instruct LLM (an LLM from IBM's Granite series)
- IBM Security & Compliance Center and Workload Protection (planned)

Watsonx data & AI platform

Watsonx.ai meets developers where they are by providing tooling that innovates across the entire AI technology stack. The platform offers a unified AI development experience with access to foundation models, a curated set of tooling including third-party frameworks such as crewAI, SDKs and APIs and templates to be successful when developing generative AI, agents or traditional Machine Learning solutions.

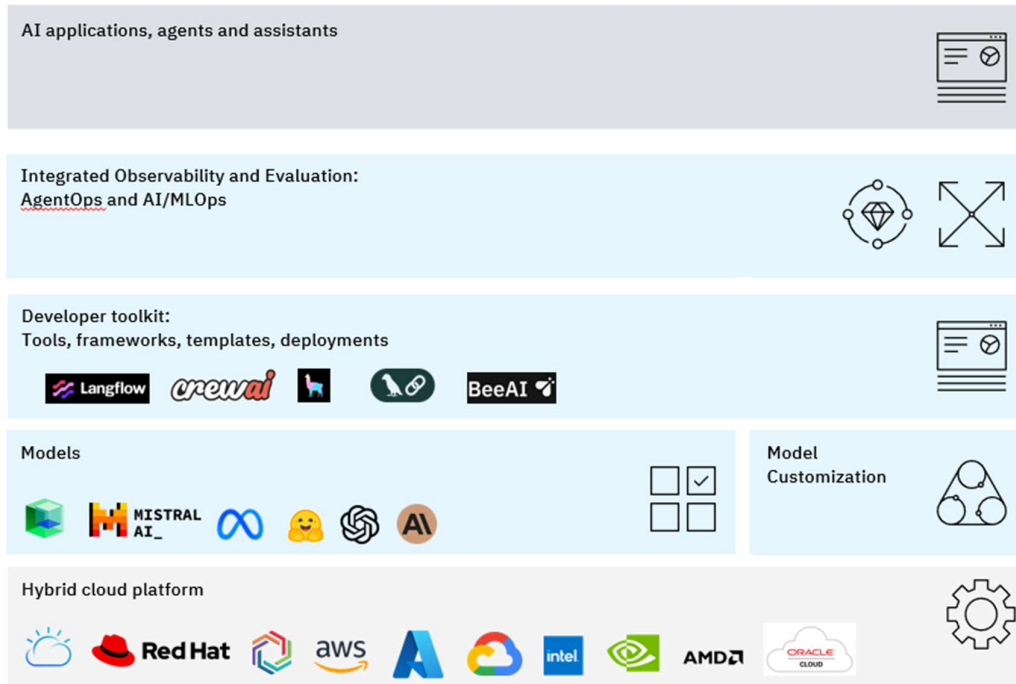


Figure 4 - AI technology stack

The product also offers extensions to your IDE of choice as well as a modern out-of-the box front-end UI for a low-code experience, surrounded by the various guardrails available within the watsonx portfolio to ensure security, privacy and governance.

Watsonx.ai runs on the OpenShift hybrid cloud platform, integrating seamlessly into existing infrastructures, systems and processes.

Watsonx.ai is modeled after the software development lifecycle, going around the flywheels above. However, it is not only for Generative AI. Watsonx.ai also has a full suite of tooling for data science and machine learning workflows that enable data science teams to build, test and deploy machine learning models for their use cases. All of this occurs in one collaborative studio.

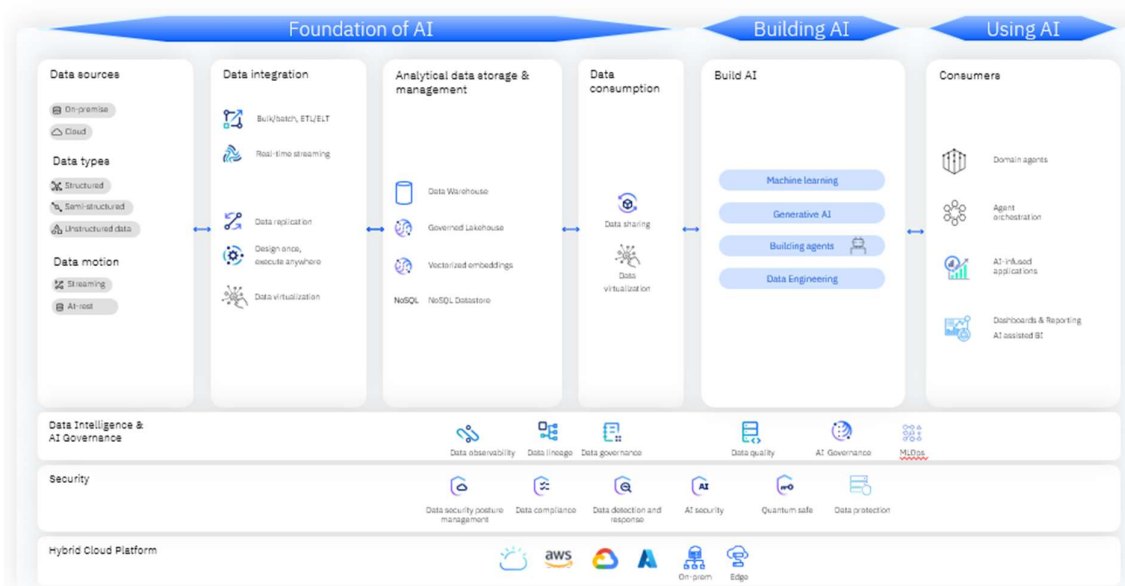


Figure 5 - IBM Data Platform Reference Architecture

Going beyond the watsonx.ai service, the IBM data & AI platform provides a large number of integrated services spanning all the way from data ingestion, transformation, data storing, building machine learning and Generative AI solutions, deployment of such solutions into development, test and production environments, and monitoring the performance of them, and ensuring not only business critical uptime SLAs are met, but also that drift, bias, fairness and other AI-specific metrics are met.

Demo video for AI Sweden DDO project (20 minutes)

This demo shows an end-to-end sample workflow from AI Governance, data governance, consuming data, building and deploying machine learning models, and monitoring in a production environment. The demo does not cover the Generative AI capabilities the platform has.

<https://www.youtube.com/watch?v=oy9luj9tq4>

See further, Simplify Your AI Journey: Unleashing the Power of AI with IBM watsonx.ai: <https://www.redbooks.ibm.com/abstracts/sg248574.html>

AI building blocks to the future

When talking about AI as the building blocks for the future it is important to start at the center of the image on the slide - use cases. What use case are businesses driving where they're applying AI first to that use case? In terms of applications, there are five different critical elements.

- Assistants
- Agents
- Models
- Data
- Governance

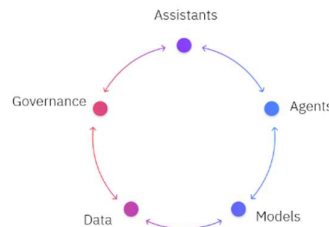


Figure 6 - AI building blocks to the future

Businesses need to ask themselves what is the AI assistant experience they provide their users to be able to interact with Generative AI, and how is that infused with more modern capabilities of Generative AI. Does the organization have Agentic AI technologies that are built on a foundation of small language models that can be customized with their data to be able to have the right kind of outcomes? Furthermore, given the use cases of the company, clients will want to be able to govern their agents and AI, whether that be with the right kind of guardrails, the risk and compliance aspects, or the AI security aspects for their business use case. This is what IBM considers as the core building blocks when thinking about how to apply AI within an organization.

AI capabilities are growing rapidly

Until the 2020s, if someone were to describe AI systems, the most common technology that would come to mind would be machine learning. This is now a mature technology, that's been widely adopted in every industry for many purposes. For example, predicting of the broadband Internet since it electrical grid loads, assessing whether a transaction is fraudulent, predictive maintenance of machines in a factory, or what additional products to recommend to an online shopper.

When the hype around Generative AI began in late 2022, pundits everywhere were proclaiming that this would mark the beginning of a new era of technical innovation, rivalling the impact became widely adopted in the late-1990s. As with the Internet, it was not immediately clear what the impact of Generative AI would be, outside of the immediately obvious. People have been generating marketing copy, application code, and movie clips. Developers have also been able to tune Generative AI

models to provide their chatbots with much more engaging conversational interfaces.

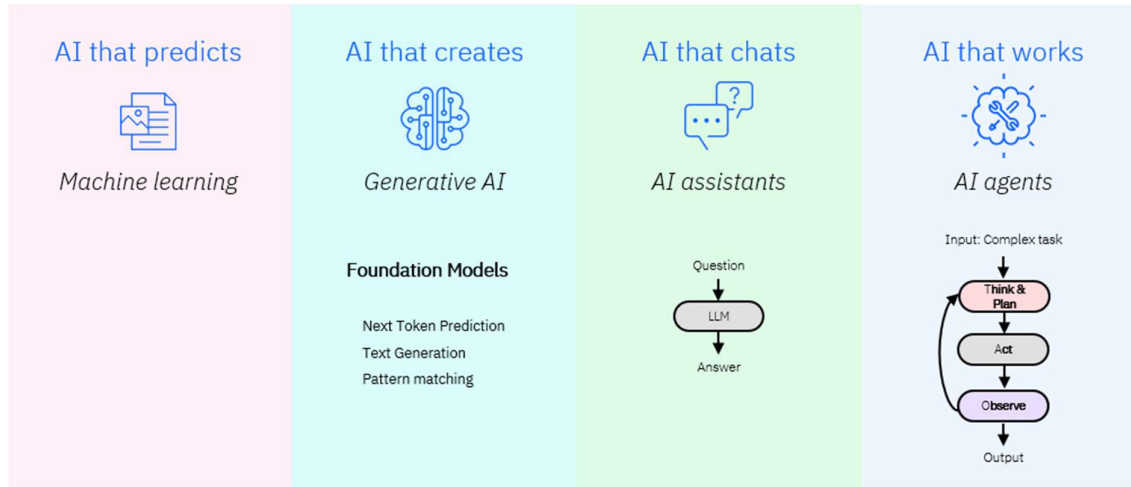


Figure 7 - AI capabilities are growing rapidly

Now, roughly three years since people were first trying ChatGPT (with the now seemingly ancient GPT 3.5 model), the truly transformative potential of Generative AI is coming into focus. In short, Generative AI has evolved so that it's not just about simply generating content but doing actual work. Just like Netflix and Uber represent game-changing killer apps on the Internet, networks of AI agents will form Agentic AI architectures, which are going to revolutionize knowledge worker productivity. The business world is getting close to a point where knowledge workers will be using – and building – AI agents, which will act as companion digital workers that complete tasks and do work on their behalf.

IBM Model strategy

IBM has always had a different and differentiating approach to AI foundation models: there will be no monopoly and no single model (no matter how large or how many different modalities it can handle). Businesses need different models to provide the best cost performance for different use cases.

IBM offers a large set of models from different vendors and, includes both open and proprietary models, in the watsonx platform. There are models specific to different tasks and languages that clients can use. The watsonx.ai platform allows clients to easily test and switch models. As it turns out, smaller models are less costly to inference and, when properly selected, prompted, tuned, and focused, can often perform as well as larger ones. Using watsonx.ai's flexible hybrid approach, models can be deployed anywhere, minimizing cost and maximizing access. The breadth of open-source models prevents clients from getting locked into any vendor or model.

See further, IBM AI Models: <https://www.ibm.com/solutions/ai-models>

See further, How to choose the right AI foundation model, What are large language models (LLMs): <https://www.ibm.com/account/reg/us-en/signup?formid=urx-52620>

See further, Simplify Your AI Journey Hybrid Open Data Lakehouse with IBM watsonx.data: <https://www.redbooks.ibm.com/redbooks/pdfs/sg248570.pdf>

See further, IBM guide for developers: <https://www.ibm.com/downloads/documents/us-en/1443d5cdd6c02c76>

AI Governance

IBM AI Governance capabilities (delivered through watsonx.governance product) was designed to direct, manage and monitor the AI activities of an organization. The solution is designed to meet regulatory requirements and ethical concerns through software automation. It drives a complete AI governance solution without the excessive costs of switching from a client's current data science platform. This solution includes processes that trace and document the origin of data, models and associated metadata and pipelines for audits. The documentation should include the techniques that trained each model, the hyperparameters used, and the metrics from testing phases. The result of this documentation is increased transparency into the model's behavior throughout the lifecycle, the data that was influential in its development, and its possible risks.

Before a model is put into production, it is validated to assess the risks to the business. Once the model goes live, it is continuously monitored for fairness, quality, drift etc. Regulators and auditors are given access to its documentation which provide explanations of the model's behaviour and predictions. These explanations provide visibility into how the model works and what processes and training the model received.

The first challenge is how to operationalize AI with confidence. Why is this a challenge? Let's take an example. One of the clients that we worked with had over 700 models with no idea of how they were built, what stage they were in, and they had no automated way to see how they were performing. The landscape was fragmented with data scientists building models using their own tools of choice. Because there was no visibility into the data used, model processes, no cataloguing or monitoring decision could not be made fast enough to move models into

production. Then there is the issue of data scientists leaving the company and not providing data and model tracking.

When we think of lack of model transparency and explainability we usually focus on models. But transparency is critical through the entire lifecycle, from acquiring the right data, to the building and testing the right model. This includes tracking to determine both bias in data and models.

The third challenge is lack of automation, without this there is no scaling. Think about the previous example of the organization deploying 700 models. Without automation it is possible to monitor and track perhaps a few models. But as you get more data, more models and more applications you quickly lose the ability to safely scale. Manual processes also introduce human errors and delay the time it takes to deploy models. During this time there can be introduction of bias and drift.

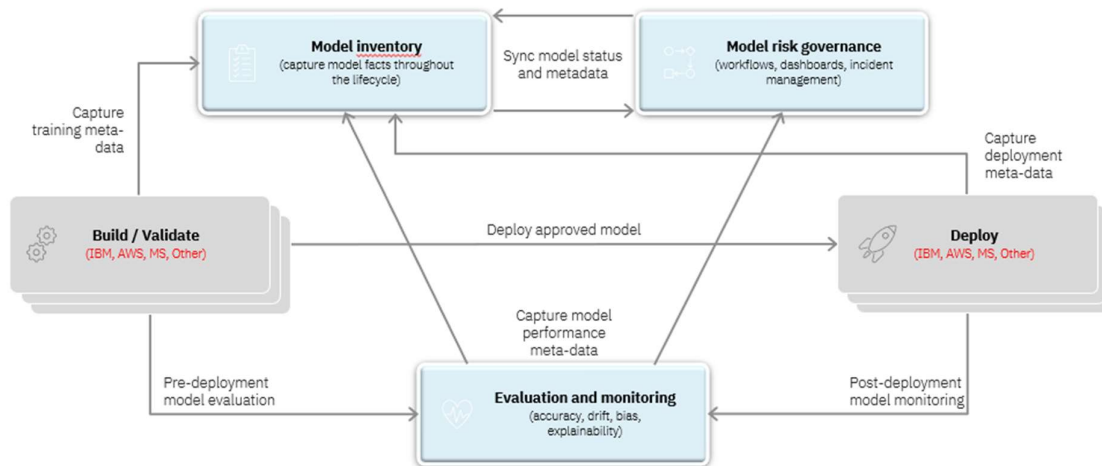


Figure 8 - IBM watsonx.governance integrations

IBM AI Governance provides three main capabilities (in blue) that work together with different AI stacks. As an example of an end-to-end governed process:

1. Once the model proposal has gone through the appropriate approval process, a model entry is created in the Model Inventory. The entry will be continuously updated with new information.
2. The data scientist uses the tool of their choice to develop the model. Training data and metrics from a number of popular open-source frameworks are automatically captured and saved to the model entry. Custom information can also be saved.
3. When the pre-production model is evaluated for accuracy, drift and bias, the performance metadata is captured and synchronized.

4. The model is reviewed and approved for production workloads.
5. In the preferred platform, the model is deployed and once again the relevant meta-data is captured and synchronized.
6. Lastly, the production model is continuously monitored, and the performance data captured and synchronized as well.

A dashboard provides a comprehensive view of the performance metrics for all models, allowing stakeholders to proactively identify and react to any issues.

See further, Simplify Your AI Journey: Ensuring Trustworthy AI with IBM watsonx.governance: <https://www.redbooks.ibm.com/abstracts/sg248573.html>

IBM Cloud

IBM Cloud is a focused, trusted Enterprise Cloud Platform, optimized for large enterprises looking for a secure and compliant landing zone for mission critical workloads, especially in regulated industries. Aligned to the heritage of IBM, we are a trusted partner differentiated on our client service proposition, relationships, and partnerships, as well as market leading security, compliance, performance, and resiliency.

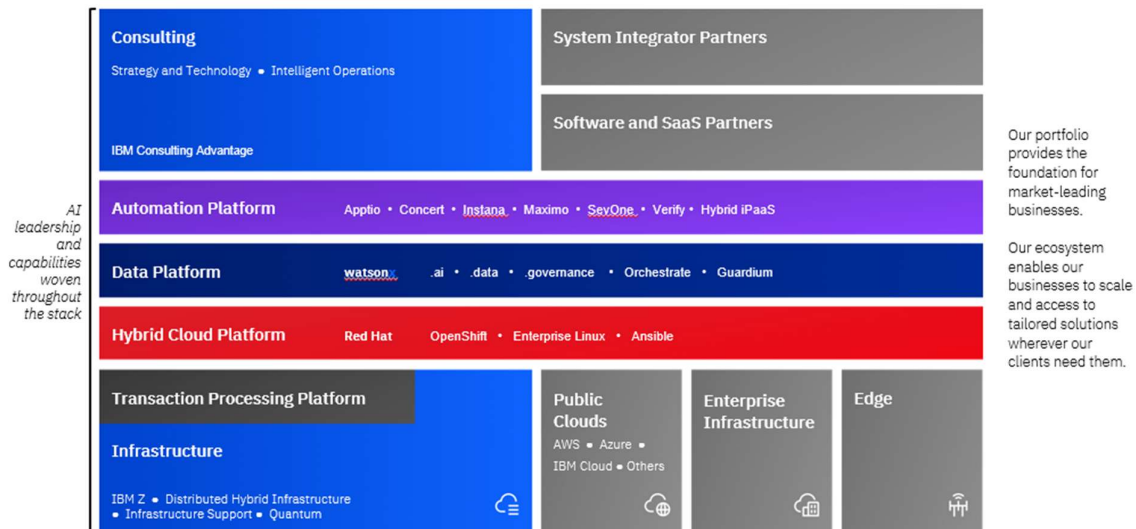


Figure 9 - IBM's Integrated Portfolio & Strategy

We address clients with incumbent IBM workloads who are looking to take advantage of cloud, protecting and growing that incumbency. Integral to IBM's Hybrid Cloud and AI strategy, our platform provides commercial and technical advantages for IBM heritage products like Db2, WAS, MQ, and IBM Power as a

Service and is the primary cloud destination for innovation such as watsonx, Red Hat OpenShift AI, and IBM Quantum computing.

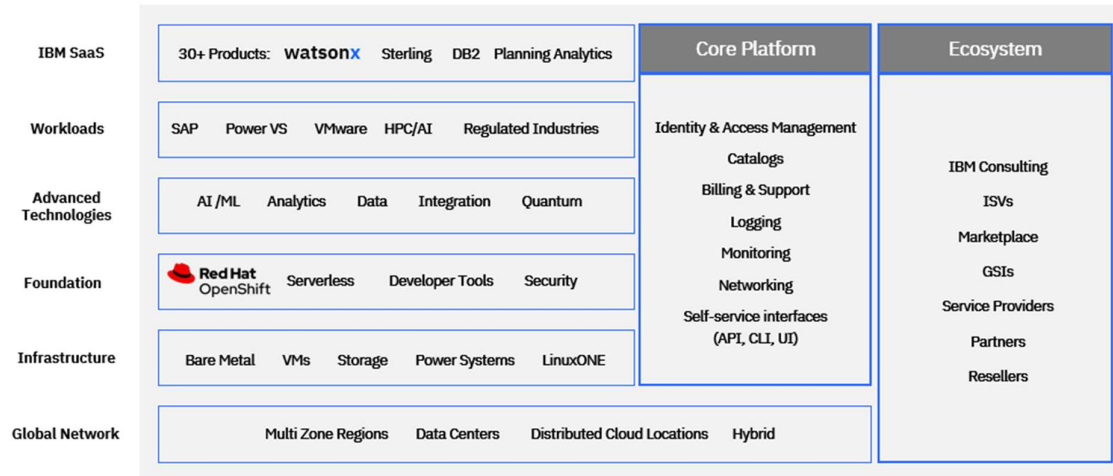


Figure 10 - IBM Cloud stack enables hybrid cloud and AI

With our open-source technologies, such as Kubernetes, Red Hat OpenShift, and a full range of compute options, including virtual machines, containers, bare metal, and serverless, you have the control and flexibility that's required to support workloads in your hybrid environment. You can deploy cloud-native apps while also ensuring workload portability.

From infrastructure to runtime services, clients can easily move their AI business models from experimentation to monetization

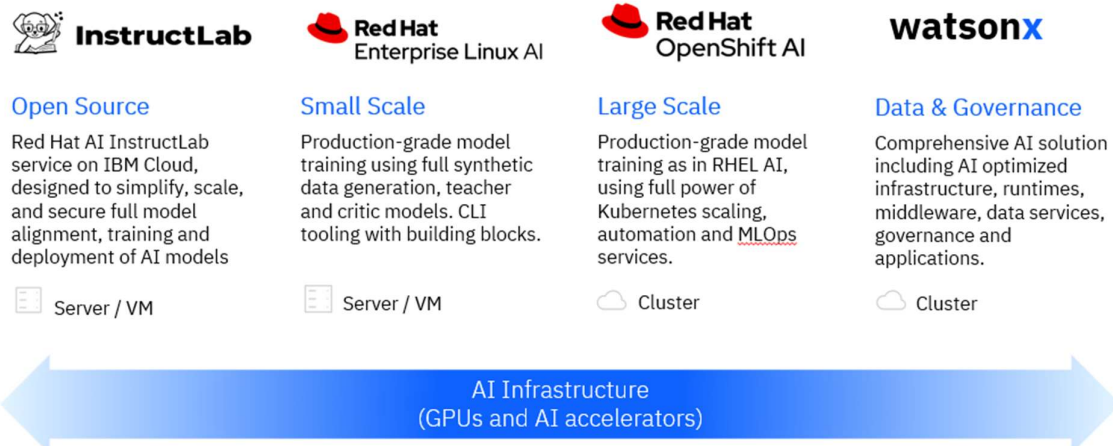


Figure 11 - IBM Cloud AI portfolio

Whether you need to migrate apps to the cloud, modernize your existing apps by using cloud services, ensure data resiliency against regional failure, or use new

paradigms and deployment topologies to innovate and build your cloud-native apps, the platform's open architecture is built to accommodate your use case.

AI Infrastructure

A hybrid cloud infrastructure allows you to effectively manage the various stages for any AI project. AI stack allows you to automate several workflows for preparing, building, deploying and maintaining inferencing, tuning, training and data management, but not all clients need to take care of all these stages on their own, with Foundation Models you don't have to build custom models for every scenario.

The reality is most clients will start from the top:

- Many clients will use assistants, perhaps the simplest approach to be in market fast.
- Next, clients that need to do inferencing, anywhere, probably consuming it as a service, which is the kind of inferencing that is needed. (In many cases with watsonx and IBM generated models, we do the data aggregation, we do the build),
- Third, some clients will do model adaptation, and, for doing this, you want to be very selective because you cannot trust your private data to go anywhere (you'd rather do it on prem or you do it with a cloud provider that you trust, on a model that you trust). And for model adaptation, IBM provides a significant advantage, aligning to the business value clients need and the required technology and support they need.
- Finally, very few clients will build their own model, it's not really expected they do distributed training.

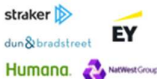



	I want	Solution	Sample use cases	
1	Out of the box generative AI solutions	Agents and AI assistants	Customer assist, agentic workflow, document processing	
2	To build customization, have model alignment, and scale AI models	Inferencing and Model-as-a-Service	Financial knowledgebase, policy/governance automation, customer satisfaction analysis	
3	To build my own AI stack, just need AI-ready infrastructure	Build Your Own software stack Deploy on resilient, secure infrastructure	LLM hosting, HPC	
4	Drive scale	Surround cloud services Security tools, observability, application platforms	Logging, monitoring, security and compliance, ISVs, cloud native	

Figure 12 - Most common patterns for AI

So, bringing IBM AI and IBM Cloud together offers unique benefits to enterprise clients and those concerned with some of the continuing challenges of adopting AI in business.

Both IBM watsonx and IBM Cloud were built with enterprise and regulated industries in mind. Watsonx emphasizes orchestration and governance while IBM Cloud adds many factors to support AI use in regulated industries with stringent compliance and security requirements.

Everything we do is built around trust, whether it is the AI software, the Gen AI models, or the infrastructure used to securely train, fine-tune, or run enterprise AI. IBM also provides open architecture supporting multi-cloud through Red Hat OpenShift.

Unique enterprise data management capabilities within IBM Cloud include the ability for clients to keep their own keys (KYOK) and industry-leading data encryption. This provides clients with peace of mind when it comes to their proprietary data as they are the only ones who can physically access their own data. Even IBM is unable to do so.

Below are recent IBM Cloud & AI Infrastructure announcements:

Announcements from IBM:	Links:
InstructLab Democratizing LLM Development	https://github.com/instruct-lab
IBM Granite Open sourced secure LLM models	https://huggingface.co/ibm-granite/granite-7b-base
RHEL AI on IBM Cloud Bootable foundation model platform	https://www.redhat.com/en/about/press-releases/red-hat-enterprise-linux-ai-now-generally-available-enterprise-ai-innovation-production
OpenShift AI Expand and scale across distributed cluster environments	https://github.com/IBM/roks-openshift-ai-da/blob/main/README.md
NVIDIA H200 Up to 2x inference performance over the H100	https://newsroom.ibm.com/blog-ibm-brings-enhanced-performance-and-efficiency-for-ai-and-hpc-with-nvidia-accelerated-computing
INTEL Gaudi 3 Deploy Gaudi 3 for inferencing on IBM Cloud and lower your TCO of your AI stack	https://newsroom.ibm.com/blog-intel-and-ibm-announce-the-availability-of-intel-gaudi-3-ai-accelerators-on-ibm-cloud
AMD MI300X Securely deploy Inferencing and HPC workloads on IBM Cloud	https://www.ibm.com/products/gpu-ai-accelerator/amd

Table links 13 - IBM Cloud AI Infrastructure

IBM Security Compliance Center and Workload Protection

Security and Compliance Center Workload Protection help you accelerate your Kubernetes and cloud adoption by addressing security and regulatory compliance. Easily identify vulnerabilities, check compliance, block threats and respond faster at every stage of the container and Kubernetes lifecycle.



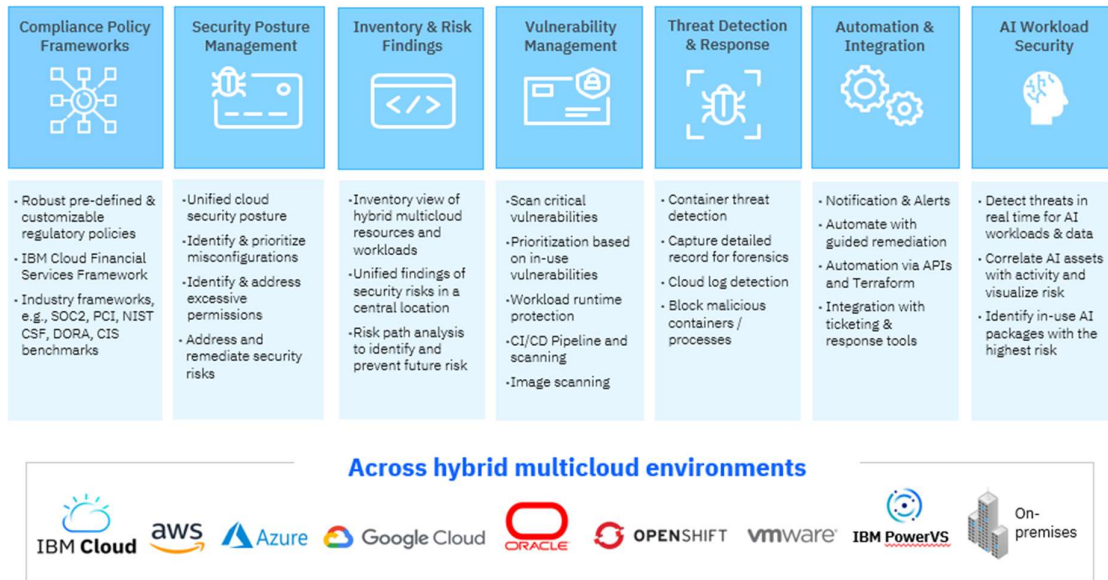


Figure 14 - Address Security, Risk and Compliance with SCC WP

Features and capabilities

Continuously validate compliance - Meet regulatory compliance standards for containers and cloud. Save time with out-of-the-box policies and reports for PCI, NIST, SOC2, etc. that map to specific controls and implement File Integrity Monitoring (FIM).

Prioritize vulnerabilities - Automate CI/CD pipeline and registry scanning without images leaving your environment. Block vulnerabilities in pre-production and monitor for new CVEs at runtime for containers and hosts. Map critical vulnerabilities back to an application and dev team.

Detect and respond to runtime threats - Secure containers, Kubernetes, OpenShift, hosts and cloud infrastructure with out-of-the-box policies based on open source Falco. Prevent lateral movement using Kubernetes network policies.

Container Forensics & Incident Response - Incident response and container forensics for Kubernetes and OpenShift. Conduct forensics and incident response to understand security breaches, meet compliance requirements and recover quickly even after a container is gone.

Cloud Native Network Security - Support a Zero Trust approach to container network security by allowing only required communication. Visualize all network communication between pods, services, and applications inside Kubernetes.

See further, <https://www.ibm.com/products/security-and-compliance-center>

IBM Client Engineering

As part of the IBM partnership with AI Sweden there are possibilities to reach out and use the IBM Client Engineering.

IBM Client Engineering delivers scalable business outcomes with a skilled, multidisciplinary squad and a human-centered approach. They meet you in your digital transformation journey, co-creating pilots to prove value. By using IBM technology and methodologies, they invest in you to drive innovation and deliver tailored solutions.

See further: <https://www.ibm.com/client-engineering>

DDO use case mappings

We have illustrated on a high level which IBM AI capabilities deployed in the AI Sweden lab (described in separate document) could be a good fit for the different DDO uses cases. Showing the potential tool, infrastructure resource, AI capability etc.

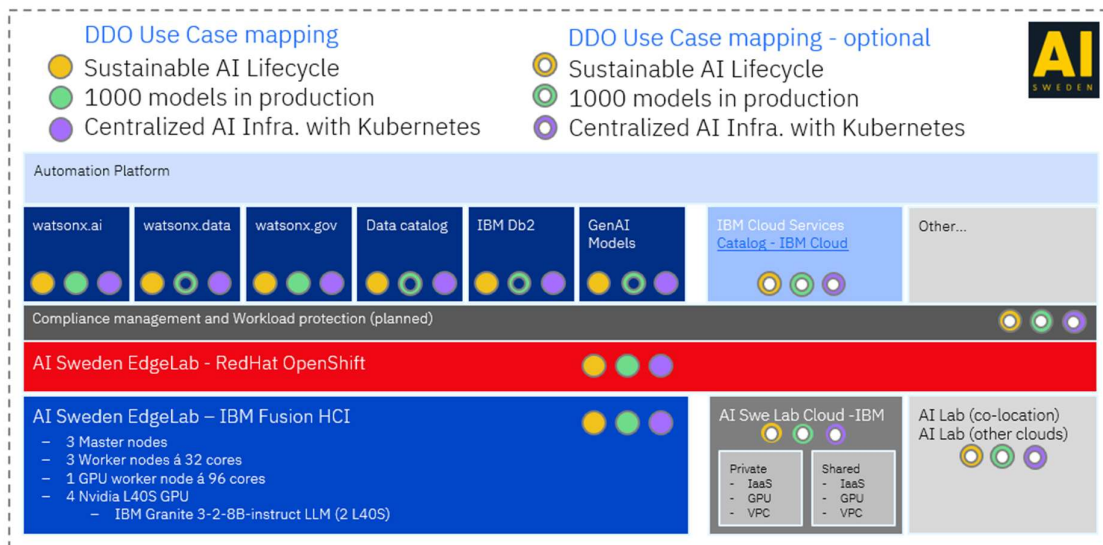


Figure 15 - IBM AI lab Fusion HCI platform with DDO use case mapping

Separate document, (IBM complement to DDO project and white papers.doc, found here, <https://www.ai.se/en/data-driven-organizations>) is describing the IBM components installed at the AI Sweden Gothenburg lab.

DDO – Best practices for operationalization of AI in Sweden

This material has been produced as part of the Vinnova-funded project *Data-driven organizations – Best practices for operationalization of AI in Sweden* (DDO), a project lasting just under two years with twenty participants from private sector, public sector, and academia. Together, they tackled issues concerning large- and small-scale operation of AI solutions and how to enable and use AI broadly across an organization.

The work focused on three specific use cases: Local sustainable operation of AI, legal and technical prerequisites for effective infrastructure, and how to create the best conditions for keeping thousands of AI models in operation.

A compilation of all material produced within the framework of DDO is available on [AI Sweden's website](#).

The organizations that participated in DDO were [Aixia](#), [Hewlett Packard Enterprise](#), [Hopsworks](#), [IBM](#), [Linköping University](#), [NetApp](#), [Predli](#), [Proact](#), [RISE](#), [RedHat](#), [Region Halland](#), [Sahlgrenska University Hospital](#), [Statistics Sweden](#) (Statistiska Centralbyrån), [The Swedish Tax Agency](#) (Skatteverket), [Stormgrid](#), [The Swedish Transport Administration](#) (Trafikverket), [Volvo Parts](#), [Region Västra Götaland](#), [Santa Anna](#), and [AI Sweden](#).

The project was funded by the participating organizations and Vinnova. AI Sweden is in part financed by the EU.

IBM watsonx presentation and demo

Henrik Sjöstrand's presentation and demo, from 2025-09-19 is available at Youtube here:

DEMO	https://youtu.be/oyu9luj9tq4
PRESENTATION	https://youtu.be/9XK1wZrizic

IBM Contacts & Contribution

The following people from IBM have been participating in writing this document:

Name :	Role :	Mail :	Phone :
John Magnusson	Brand Technical Specialist, Cloud	john.magnusson@se.ibm.com	070-7931677
Henrik Sjöstrand	Senior Partner Technical Specialist, Data & AI	henrik.sjostrand@ibm.com	070-7935556
Roger Eriksson	Senior Partner Technical Specialist, Storage	roger_eriksson@se.ibm.com	070-7933518
Johan Rodin	Advisory Partner Technical Specialist - Data&AI	johan.rodin@se.ibm.com	070-7931807
Wanmeng He	AI Accelerator Leader, Client Engineering NCEE	Wanmeng.He@ibm.com	070-2929844
Holger Hellebro	Executive Architect and Data Scientist, AI & Analytics	holger.hellebro@se.ibm.com	070-7935715
Håkan Andersson	Technical Community Leader & CTO	hakan.c.andersson@se.ibm.com	070-7931082